



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2016-0066387
(43) 공개일자 2016년06월10일

(51) 국제특허분류(Int. Cl.)
G06F 17/27 (2006.01) G06F 17/28 (2006.01)
(21) 출원번호 10-2014-0170717
(22) 출원일자 2014년12월02일
심사청구일자 2014년12월02일
기술이전 희망 : 기술양도

(71) 출원인
울산대학교 산학협력단
울산광역시 남구 대학로 93
(72) 발명자
옥철영
울산광역시 남구 신북로35번길 6, 101동 602호 (무거동, 쌍용스윗닷홈)
신준철
울산광역시 남구 대학로47번길 6, 2층 B1호 (무거동)
(74) 대리인
김중선, 이형석

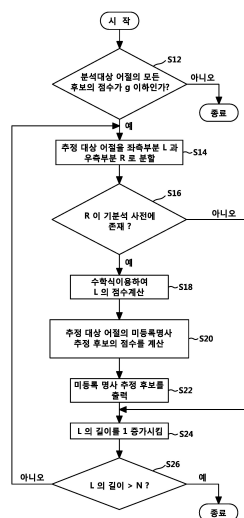
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 형태소 분석기를 위한 미등록 명사 추정 장치 및 방법

(57) 요약

본 발명은 형태소 분석기를 위한 미등록 명사 추정 장치 및 방법에 관한 것으로, 상세하게는, 미리 형태소들을 등록하고 등록된 정보를 이용하여 형태소 분석을 수행할 때 등록되지 않은 미등록어의 경우에는 정상적으로 분석하지 못하는 문제점을 해결하기 위하여, 미등록 명사가 포함된 어절의 경우에 좌측부분L과 우측부분R로 나누고 우측부분R이 기분석 사전에 존재하면 좌측부분의 길이, 우측부분의 길이 및 제안된 수확식을 이용하여 좌측부분의 빈도 점수를 계산하며, 기분석 사전에서 검색된 우측부분이 빈도, 계산된 좌측부분의 빈도 점수, 및 좌측부분과 우측부분 간 조합 가능성 정도를 곱하여 후보 점수를 계산함으로써 가장 높은 점수의 미등록 명사 추정 후보를 미등록 명사로 추정하는 형태소 분석기를 위한 미등록 명사 추정 장치 및 방법을 개시한다.

대표도 - 도2



이 발명을 지원한 국가연구개발사업

과제고유번호 1711006603

부처명 미래창조과학부

연구관리전문기관 한국산업기술평가관리원

연구사업명 SW컴퓨팅산업원천기술개발

연구과제명 비기호적 기법 기반 인간모사형 자가학습 지능 원천기술 개발

기 여 율 1/2

주관기관 포항공과대학교산학협력단

연구기간 2014.05.01 ~ 2015.02.28이 발명을 지원한 국가연구개발사업

과제고유번호 2014R1A1A2009506

부처명 교육과학기술부

연구관리전문기관 한국연구재단

연구사업명 일반연구자지원사업(리서치/지역대학)

연구과제명 기계학습 기반 형태소 분석기의 도메인 최적화와 개체명 인식 기능

기 여 율 1/2

주관기관 울산대학교산학협력단

연구기간 2014.06.01 ~ 2015.05.31

명세서

청구범위

청구항 1

미리 형태소들이 등록되어 있는 기분석사전을 이용하여 형태소 분석을 수행하는 형태소 분석기에 의해서 계산된 형태소 분석 대상 어절의 모든 후보의 점수가 추정 임계치 g 이하이면, 미등록 명사가 포함된 추정 대상 어절이라고 판단하고 추정 대상 어절을 어절 길이의 경우의 수만큼 좌측부분과 우측부분으로 나누어 각 좌측부분 및 우측부분에 대하여 미등록 명사 추정 후보의 점수를 계산하며, 가장 높은 점수의 미등록 명사 추정 후보를 미등록 명사로 추정하고 추정된 미등록 명사를 수집하는 미등록 명사 추정/수집부; 및

상기 형태소 분석 대상 어절에 대해 분석된 형태소를 태깅하고 상기 추정 대상 어절에 대해 추정된 미등록 명사를 태깅하는 태깅부;를 포함하는 형태소 분석기를 위한 미등록 명사 추정 장치.

청구항 2

제 1 항에 있어서,

추정 대상 어절을 출력하는 디스플레이부; 및

상기 출력된 추정 대상 어절을 확인한 사용자로부터 미등록 명사 여부를 결정하는 결정 신호를 입력받는 입력부를 더 포함하며,

상기 미등록 명사 추정/수집부는, 상기 미등록 명사 추정 후보의 가장 높은 점수가 반자동 임계치 이하이면 상기 추정 대상 어절의 출력 및 수집을 결정하고 상기 입력부로부터 입력된 사용자의 결정 신호가 미등록 명사를 나타내는 경우에 상기 추정 대상 어절이 미등록 명사임을 결정하는, 형태소 분석기를 위한 미등록 명사 추정 장치.

청구항 3

제 1 항 또는 제 2 항에 있어서, 상기 미등록 명사 추정/수집부는,

상기 추정 대상 어절을 어절 길이의 경우의 수만큼 좌측부분과 우측부분으로 나누어 각 좌측부분 및 그의 우측부분에 대하여 미등록 명사 추정 후보의 점수를 계산하되,

좌측부분에 대응되는 우측부분이 기분석 사전에 존재하는지를 확인하고, 기분석 사전에 존재하는 우측부분의 길이, 상기 좌측부분의 길이 및 수학식을 이용하여 상기 좌측부분의 빈도 점수를 계산하며,

상기 계산된 좌측부분의 빈도 점수, 기분석 사전에 존재하는 우측부분의 빈도로부터 수정된 빈도값, 상기 좌측부분 및 상기 우측부분의 조합 가능성 정도를 곱하여 미등록 명사 추정 후보의 점수를 계산하는, 형태소 분석기를 위한 미등록 명사 추정 장치.

청구항 4

제 3항에 있어서,

상기 좌측부분의 빈도 점수를 계산하기 위한 상기 수학식은,

$$l = \frac{(a \times s_r + b)}{c^{d \times S_l + e}}$$

이고, l 은 미등록 명사 부분 점수이며, s_r 은 우측부분인 조사부분의 길이, s_l 은 좌측부분인 미등록 명사 부분의 길이, 그리고 a, b, c, d, e 는 상수이며,

상기 기분석 사전에 존재하는 우측부분의 빈도로부터 수정된 빈도값은,

$$r' = r^f$$

이고, r 은 우측부분인 조사부분의 기분석 사전에 존재하는 빈도이며, r' 는 수정된 빈도값이고, f 는 상수인 것을 특징으로 하는 형태소 분석기를 위한 미등록 명사 추정 장치.

청구항 5

형태소 분석 대상 어절의 모든 후보의 점수가 추정 임계치 g 이하이면, 형태소 분석 대상 어절이 미등록 명사가 포함된 추정 대상 어절이라고 판단하여 미등록 명사 추정을 시도하는 제 1 단계;

추정 대상 어절을 어절 길이의 경우의 수만큼 좌측부분과 우측부분으로 나누어 각 좌측부분 및 우측부분에 대하여 미등록 명사 추정 후보의 점수를 계산하는 제 2 단계; 및

가장 높은 점수의 미등록 명사 추정 후보를 미등록 명사로 추정하는 제 3 단계;를 포함하는 형태소 분석기를 위한 미등록 명사 추정 방법.

청구항 6

제 5항에 있어서,

상기 미등록 명사 추정 후보들의 가장 높은 점수가 반자동 임계치 이하이면 상기 추정 대상 어절을 출력하는 단계; 및

상기 추정 대상 어절을 확인한 사용자로부터 미등록 명사 여부를 결정하는 결정 신호를 입력받아 미등록 명사 여부를 결정하는 단계를 더 포함하는 형태소 분석기를 위한 미등록 명사 추정 방법.

청구항 7

제 5 항 또는 제 6 항에 있어서, 상기 제 2 단계는,

좌측부분에 대응되는 우측부분이 기분석 사전에 존재하는지를 확인하는 단계;

기분석 사전에 존재하는 우측부분의 길이, 상기 좌측부분의 길이 및 수학식을 이용하여 상기 좌측부분의 빈도 점수를 계산하는 단계; 및

상기 계산된 좌측부분의 빈도 점수, 기분석 사전에 존재하는 우측부분의 빈도로부터 수정된 빈도값, 상기 좌측부분 및 상기 우측부분의 조합 가능성 정도를 곱하여 미등록 명사 추정 후보의 점수를 계산하는 단계를 포함하는 형태소 분석기를 위한 미등록 명사 추정 방법.

청구항 8

제 7 항에 있어서,

상기 좌측부분의 빈도 점수를 계산하기 위한 상기 수학식은,

상기 우측부분의 길이가 길수록 점수가 높아지고 상기 좌측부분의 길이가 길수록 점수가 낮아지게 계산되도록 형성되며,

상기 기분석 사전에 존재하는 우측부분의 빈도로부터 수정된 빈도값은,

상기 기분석 사전에 존재하는 조사 부분의 길이가 짧을수록 후보 점수가 높아지는 성향을, 길이가 길수록 점수

가 높아지도록 보상하기 위하여 수정된 값인 것을 특징으로 하는 형태소 분석기를 위한 미등록 명사 추정 방법.

청구항 9

제 7 항에 있어서,

상기 좌측부분의 빈도 점수를 계산하기 위한 상기 수학식은,

$$l = \frac{(a \times s_r + b)}{c^{d \times s_l + e}}$$

이고, l은 미등록 명사 부분 점수이며, s_r은 우측부분인 조사부분의 길이, s_l은 좌측부분인 미등록 명사 부분의 길이, 그리고 a,b,c,d,e는 상수이며,

상기 기분석 사전에 존재하는 우측부분의 빈도로부터 수정된 빈도값은,

$$r' = r^f$$

이고, r은 우측부분인 조사부분의 기분석 사전에 존재하는 빈도이며, r'는 수정된 빈도값이고, f는 상수인 것을 특징으로 하는 형태소 분석기를 위한 미등록 명사 추정 방법.

청구항 10

제 5 항 또는 제 6 항에 있어서, 상기 제 2 단계는,

좌측부분의 길이가 1개인 경우부터 어절 전체 길이인 경우까지 순차적으로 반복 수행되는 것을 특징으로 하는 형태소 분석기를 위한 미등록 명사 추정 방법.

발명의 설명

기술 분야

[0001] 본 발명은 형태소 분석기를 위한 미등록 명사 추정 장치 및 방법에 관한 것으로, 더욱 상세하게는, 미리 형태소들을 등록하고 등록된 정보를 이용하여 형태소 분석을 수행할 때 등록되지 않은 미등록어의 경우에는 미등록어의 품사를 추정하거나 형태소 경계를 추정하여 미등록 명사를 인식하도록 함으로써 형태소 분석기의 정확률을 향상시키는 형태소 분석기를 위한 미등록 명사 추정 장치 및 방법에 관한 것이다.

배경 기술

[0002] 일반적으로 한국어 처리를 위해서는 형태소 분석기가 필요하기 때문에 다양한 형태소 분석기가 연구되고 있다. 한국어는 하나의 어절에 여러 형태소가 조합될 수 있는 교착어이기 때문에 자동 형태소 분석에 많은 어려움이 있다. 대표적으로 어려운 점으로는 조사와 명사의 경계를 구분하는 것과, 복합명사와 미등록 명사를 구분하여 처리하는 것이다. 이 문제들은 모두 동시에 발생할 수 있기 때문에 하나의 문제로 인식할 수도 있다.

[0003] 대체로 형태소 분석기들은 미리 형태소들을 등록하고 이 정보를 적극적으로 이용하여 형태소 분석을 한다. 이런 특징 때문에 만약에 등록되지 않은 형태소를 만나게 되면 정확률이 급격히 낮아지는 문제점이 있다. 대표적으로 발생하는 오류는 1음절 또는 2~3음절의 명사들로 구성된 복합명사로 인식하는 것이며, 가끔 조사와 명사의 경계를 구분하지 못하기도 한다. 이러한 문제를 해결하기 위해서는 미등록 명사임을 추정하여 인식하는 기술이 필요하다.

[0004] 종래기술(박재한,김명선,노대욱,나대열, "백오프 통계정보를 이용한 미등록어 포함 복합명사의 분해", 제16회 한글 및 한국어 정보처리 학술대회, 제16권, 제1호, pp.65-72, 2004)에 따르면, 일반적인 복합명사와 미등록 외래어를 포함한 복합명사를 잘 분해하기 위해서 1,000만 어절의 세종말뭉치에서 448만개 명사와 복합명사를 분리

해 놓은 것에서 백오프 통계 정보를 학습하여 사용하는 연구를 수행하였다. 통계 정보는 음절 바이그램, 어휘 바이그램, 품사 바이그램 등으로 구성된다. 실험 결과 이와 같은 기계학습 방식을 사용한 미등록 명사 추정 모델에 의해 형태소 분석기 전체 시스템의 성능이 향상되는 것을 확인하였다.

- [0005] 그러나, 이와 같은 기계학습 방식을 사용한 미등록 명사 추정 알고리즘은 비교적 정확률이 높지만 학습 과정을 필요로 하고 그 결과물인 학습 정보를 사용해야만 작동이 가능하다는 단점이 있다. 이에 종래의 형태소 분석기와 함께 작동하기 위해서는 학습 시간과 추가적인 학습 정보 및 그에 따른 메모리 사용 문제에 부담이 되는 단점이 있다.

선행기술문헌

특허문헌

- [0006] (특허문헌 0001) 한국등록특허 10-0757340 호 (2007. 09. 04)
(특허문헌 0002) 한국등록특허 10-0835706 호 (2008. 05. 30)

발명의 내용

해결하려는 과제

- [0007] 본 발명은 상기한 종래 기술의 문제점을 해결하기 위하여 제안된 것으로, 본 발명의 목적은 미리 형태소들이 등록되어 있는 기본식사전을 통하여 형태소 분석을 수행하는 형태소 분석기를 이용하되, 등록되지 않은 미등록어의 경우에 미등록어를 좌측부분L과 우측부분R의 두 부분으로 분할하고 기본식사전을 검색하여 우측부분R의 빈도를 획득하고 좌측부분L의 빈도 점수를 계산하며 우측부분R의 빈도, 좌측부분L의 빈도 점수 및 이 둘의 조합 가능성 정도를 모두 곱하여 후보 점수를 계산하고 가장 높은 점수의 후보를 미등록 명사로 추정함으로써 미등록 명사를 인식하도록 하여 형태소 분석기의 정확률을 향상시키는, 형태소 분석기를 위한 미등록 명사 추정 장치 및 방법을 제공하는 것이다.

- [0008] 본 발명의 다른 목적은 추정된 후보 점수가 임계치 이하이면 디스플레이함으로써 사용자가 디스플레이된 후보 미등록어를 직접 확인하여 고유명사인지 복합명사인지를 구분할 수 있도록 반자동으로 수행하여 미등록 명사 추정의 정확성을 높이는, 형태소 분석기를 위한 미등록 명사 추정 장치 및 방법을 제공하는 것이다.

과제의 해결 수단

- [0009] 상기한 목적을 달성하기 위하여 본 발명의 일 실시예에 의한 형태소 분석기를 위한 미등록 명사 추정 장치는, 미리 형태소들이 등록되어 있는 기본식사전을 이용하여 형태소 분석을 수행하는 형태소 분석기에 의해서 계산된 형태소 분석 대상 어절의 모든 후보의 점수가 추정 임계치 g 이하이면, 미등록 명사가 포함된 추정 대상 어절이라고 판단하고 추정 대상 어절을 어절 길이의 경우의 수만큼 좌측부분과 우측부분으로 나누어 각 좌측부분 및 우측부분에 대하여 미등록 명사 추정 후보의 점수를 계산하며, 가장 높은 점수의 미등록 명사 추정 후보를 미등록 명사로 추정하고 추정된 미등록 명사를 수집하는 미등록 명사 추정/수집부; 및 상기 형태소 분석 대상 어절에 대해 분석된 형태소를 태깅하고 상기 추정 대상 어절에 대해 추정된 미등록 명사를 태깅하는 태깅부;를 포함한다.

- [0010] 상기 형태소 분석기를 위한 미등록 명사 추정 장치는, 추정 대상 어절을 출력하는 디스플레이부; 및 상기 출력된 추정 대상 어절을 확인한 사용자로부터 미등록 명사 여부를 결정하는 결정 신호를 입력받는 입력부를 더 포함하며, 상기 미등록 명사 추정/수집부는, 상기 미등록 명사 추정 후보의 가장 높은 점수가 반자동 임계치 이하이면 상기 추정 대상 어절의 출력 및 수집을 결정하고 상기 입력부로부터 입력된 사용자의 결정 신호가 미등록 명사를 나타내는 경우에 상기 추정 대상 어절이 미등록 명사임을 결정한다.

- [0011] 상기한 목적을 달성하기 위하여 본 발명의 다른 실시예에 의한 형태소 분석기를 위한 미등록 명사 추정 방법은, 형태소 분석 대상 어절의 모든 후보의 점수가 추정 임계치 g 이하이면, 형태소 분석 대상 어절이 미등록 명사가

포함된 추정 대상 어절이라고 판단하여 미등록 명사 추정을 시도하는 제 1 단계; 추정 대상 어절을 어절 길이의 경우의 수만큼 좌측부분과 우측부분으로 나누어 각 좌측부분 및 우측부분에 대하여 미등록 명사 추정 후보의 점수를 계산하는 제 2 단계; 및 가장 높은 점수의 미등록 명사 추정 후보를 미등록 명사로 추정하는 제 3 단계;를 포함한다.

[0012] 상기 형태소 분석기를 위한 미등록 명사 추정 방법은, 상기 미등록 명사 추정 후보들의 가장 높은 점수가 반자동 임계치 이하이면 상기 추정 대상 어절을 출력하는 단계; 및 상기 추정 대상 어절을 확인한 사용자로부터 미등록 명사 여부를 결정하는 결정 신호를 입력받아 미등록 명사 여부를 결정하는 단계를 더 포함한다.

발명의 효과

[0013] 본 발명은 미등록 명사 추정을 위한 추가적인 학습 및 그로 인한 추가적인 학습 정보를 필요로 하지 않고 최근에 연구된 형태소 분석기의 내부 학습사전만을 이용하므로 구현이 용이하고 컴퓨터 자원의 사용을 크게 늘리지 않으면서도 형태소 분석기의 전체 성능을 향상시킬 수 있다.

[0014] 본 발명은 기존의 형태소 분석기 알고리즘을 최대한 활용하면서 미등록 명사 추정을 정확하게 수행할 수 있고, 품사 결정 및 형태소 경계를 결정할 수 있다.

[0015] 아울러 본 발명은 추정된 미등록 명사 후보 점수가 임계값 이하이면 사용자에게 디스플레이함으로써 사용자가 디스플레이된 후보 미등록어를 직접 확인하여 고유명사인지 복합명사인지 구분하여 결정 신호를 입력할 수 있도록 반자동으로 이루어져 미등록 명사 추정의 정확성을 현저히 높이는 효과가 있다.

도면의 간단한 설명

[0016] 도 1은 본 발명의 일 실시예에 의한 형태소 분석기를 위한 미등록 명사 추정 장치의 블록 구성도

도 2는 본 발명의 다른 실시예에 의한 형태소 분석기를 위한 미등록 명사 추정 방법의 흐름도

발명을 실시하기 위한 구체적인 내용

[0017] 본 발명에 따르면, 제안하는 미등록 명사 추정 및 수집 방법을 알고리즘 모듈로 설계하여 기존의 형태소 분석기 알고리즘의 일부로 구성함으로써 별도의 학습 시간과 학습 결과물인 학습 정보를 필요로 하지 않으면서도 구현이 용이하고 속도가 빠르며 형태소 분석기의 정확도를 향상시킨다.

[0018] 본 발명은 인덱싱, 문서분류, 스팸메일 필터링 등의 주업무를 수행하기 전에 선처리에 적용되는 것으로서, 전체 문서를 스캔하고 미등록어를 처리하는 과정에 적용되는 것이다. 즉, 자연어 처리 이전에 사용되는 것이며, 본 발명에서는 미등록어의 형태소 경계를 구분하여 인식하는 것이다.

[0019] 이하, 본 발명의 바람직한 실시예들을 첨부된 도면을 참조하여 상세히 설명하기로 한다.

[0020] 본 발명에 의한 미등록 명사 추정 알고리즘은 기존 형태소 분석기와 함께 사용될 것이기 때문에 그 알고리즘이 적용될 형태소 분석기를 파악하고 어울리도록 설계하는 것이 중요하다.

[0021] 본 발명이 적용될 형태소 분석기는 기분석 부분어절 사전을 활용한 분석기에 부분어절 조건부 확률 기반의 태깅 모델을 조합한 UTagger를 사용한다. 이 형태소 분석기는 우선 1,100만 어절의 세종말뭉치를 학습하여 기분석 사전을 구축하고 어절을 분석할 때에 기분석 사전을 활용한다. 여기서 사용하는 기분석 사전은 말뭉치에 나타난 어절의 분석된 형태를 저장하고 있으며, 어절의 부분도 그에 해당하는 분석 정보와 함께 저장한다. 예를 들어 '사과를'의 기분석 내용은 <표 1>과 같다.

표 1

[0022]

분석내용	빈도	형태
사과_05[apple]/NNG+를/JKO	90	전체
사과_08[apology]/NNG+를/JKO	88	전체
사과_05[apple]/NNG+를/JKO	3	부분

사과_08[apology]/NNG+를/JKO	2	부분
--------------------------	---	----

[0023] [표 1]에서 '사과를'이 어떤 어절의 부분으로 나타난 경우는 총 5번으로 3과 2를 더한 것이다. 여기서 '사과'의 의미가 apple인 것은 총 93개로 90과 3을 더한 것이다.

[0024] [표 2]는 '를'의 기본식 내용이다. '를'은 어절의 전체로 등장한 경우보다 부분으로 등장한 경우가 많은 것을 확인할 수 있다.

표 2

분석 내용	빈도	형태
를/JKO	275,963	부분
를/JKO	11,729	전체

[0026] UTagger는 만약 분석 대상 어절 전체가 기본식 사전에 등록되어 있지 않다면 어절을 두 개로 나누어서 기본식 사전을 검색하고, 두 개로 나누어서도 해결하지 못했다면 복합명사의 가능성이 있는 것으로 간주하여 더 잘게 나누어 분석을 시도한다.

[0027] 어절을 두 개 이상으로 나눌 때 일반적으로 어절의 최우측 부분은 조사를 포함하는 형식 형태소이며 기본식 사전에서 발견되기 용이하다. 예를 들어서 '고급정보라면'은 세종말뭉치에 전체 어절이 등록되어 있지는 않지만, '~라면'은 다 수개가 발견될 수 있다. 만약 미등록 명사가 포함된 어절이라면 최우측 부분은 기본식 사전에서 발견되고 나머지는 발견되지 못했거나 발견되었더라도 매우 어색하거나 문법적으로 틀린 분석일 확률이 높다.

[0028] 형태소 분석기(UTagger)는 각 후보들을 생성할 때 문맥 정보를 고려하지 않고 오직 해당 어절 정보만을 고려하여 후보 점수를 계산한다. 이 점수는 비록 문맥 정보를 고려한 것은 아니지만 어절 내부의 자연스러움과 문법적인 가능성을 계산한 것이기 때문에 지나치게 낮다면 해당 후보는 태깅 단계로 넘어가기 전에 삭제될 수 있으며, 태깅 단계에서는 문맥 정보와 함께 사용되기 때문에 매우 중요하다. 만약 분석 대상 어절의 모든 후보의 점수가 매우 낮다면 그 어절은 미등록 명사를 포함하고 있다고 추정할 수 있다. 따라서 분석 대상 어절의 모든 후보의 점수가 특정 점수(즉, 추정 임계치) 이하인지 아닌지를 비교하여 미등록 명사 추정의 시도 여부를 판단한다. 즉, 분석 대상 어절의 모든 후보의 점수가 추정 임계치 이하이면 미등록 명사를 포함하고 있다고 판단하여 미등록 명사 추정을 시도한다.

[0029] 형태소 분석기(UTagger)에서 미등록 명사 추정을 한다는 것은 새로운 후보(미등록 명사 추정 후보)를 생성한다는 의미이며, 새로운 후보는 다른 후보와 마찬가지로 점수가 필요하다. 이 점수는 다른 후보들의 점수와 비교할 수 있도록 비슷한 방법으로 계산되어야 한다. 그렇게 된다면 미등록 명사 추정으로 생성된 후보의 점수와, 복합명사 추정으로 생성된 후보의 점수를 비교하여 해당 어절이 미등록 명사를 포함한 것인지 여부를 판단할 수도 있다.

[0030] 일단 미등록 명사 추정을 시도하게 되면 어절의 최우측 부분의 기본식 정보가 조사인지 확인하고 활용하는 것이 우선이다. 이것은 최장조사 분리방법과 흡사하지만 반드시 최장일치를 사용하는 것은 아니다. 본 발명에 의한 방법은 기본식 사전에서 발견된 빈도 정보와 길이 정보를 함께 사용하여 "미등록 명사 추정 후보"의 점수를 계산하는 것이다.

[0031] 본 발명이 적용되는 형태소 분석기(UTagger)는 일반적으로 어절을 좌측부분과 우측부분 두 부분으로 나누고 어절의 좌측 분석의 빈도와 우측 분석의 빈도 그리고 그 둘의 조합 가능성 정도를 모두 곱하여 후보 점수를 계산한다. 조합 가능성 정도란, 예를 들어서 명사 다음에는 조사가 올 확률이 높다는 것을 적용하기 위한 품사 바이그램 학습 정보와, 명사 끝 글자의 중성이 존재하면 조사 '을'이 다음에 올 수 없다는 등의 규칙 정보들을 사용하여 계산된 수치이다.

[0032] "복합명사 추정 후보"는 좌측 분석이 말뭉치, 즉, 기본식 사전에서 검색된 적이 없는 새로운 복합명사로 추정된 것이기 때문에 빈도가 '0'이어서 일반적인 후보 점수 계산 방법을 적용할 수 없다.

[0033] 따라서 본 발명에서는 좌측 빈도를 대체할 좌측 점수를 계산하여서 이것을 좌측 빈도로서 이용한다. 빈도는 0 또는 양의 정수이지만 추정된 복합명사의 점수는 대체로 0에서 1 사이로 계산된다. 본 발명에서 이 좌측 점수를

사용하여 후보 점수를 계산할 수 있다.

[0034] 본 발명에 의한 미등록 명사 추정 후보의 점수 계산에서도 미등록 명사로 추정되는 좌측 부분의 점수만 계산하면 형태소 분석기의 기본식 사전 정보와 조합 가능성 정도를 통해 후보 점수를 계산할 수 있다.

[0035] 결국 형태소 분석기에서 미등록 추정 기능을 구현하기 위해서 가장 중요한 것은 어절 내에서 미등록 명사로 추정되는 좌측 부분에 대한 점수를 계산하는 것이다.

[0036] 본 발명에서는 추가적인 학습과정 없이 간단하게 어절 내에서 미등록 명사로 추정되는 좌측 부분의 점수를 계산하는 방법을 제안한다. 즉, 조사부분의 길이와 미등록 명사 부분의 길이 정보만을 사용하여 미등록 명사 부분의 점수를 계산하는 것이다.

[0037] 본 발명에 의한 미등록 명사 부분(즉, 좌측 부분)의 점수 계산식은 [수식 1]과 같다. 즉, 조사부분의 길이가 길수록 점수가 높아지고, 반대로 미등록 명사 부분의 길이가 길수록 점수가 낮아지는 기법으로 계산된다. [수식 1]에서 1은 미등록 명사 부분 점수이며, s_r 은 우측부분인 조사부분의 길이, s_l 은 좌측부분인 미등록 명사 부분의 길이, 그리고 a, b, c, d, e 는 상수이다. 각 상수들은 해당 문제에 대한 기본적인 이해와 반복 실험을 거쳐서 사용자에게 의하여 직접 결정된 것이다.

수학식 1

$$l = \frac{(a \times s_r + b)}{c^{d \times s_l + e}}$$

[0038]

[0039] 또한 본 발명이 적용되는 형태소 분석기의 기본식 사전은 등록된 어절 또는 부분어절이 짧을수록 빈도가 매우 높아지는 성향을 보인다. 이 때문에 조사부분의 길이가 짧을수록 후보 점수가 높게 계산될 수 있다. 따라서 본 발명에서는 [수학식 2]와 같이 조사부분의 빈도를 수정하여 사용한다.

수학식 2

$$r' = r^f$$

[0040]

[0041] r 은 우측부분인 조사부분의 빈도이며, r' 은 수정된 빈도이다. 그리고 f 는 [수학식 1]의 $a \sim e$ 와 마찬가지로 상수이며, 조사부분의 길이가 짧을수록 후보 점수가 높게 계산되는 것을 보상하기 위하여 반복 실험을 거쳐서 사용자에게 의하여 직접 결정된 것이다.

[0042] 도 1은 본 발명의 일 실시예에 의한 형태소 분석기를 위한 미등록 명사 추정 장치의 블록 구성도를 나타낸다.

[0043] 도 1에 도시된 바와 같이, 본 발명의 일 실시예에 의한 형태소 분석기를 위한 미등록 명사 추정 장치는, 미리 형태소들이 등록되어 있는 기본식사전을 저장하는 DB(30)와, 상기 DB(30)에 저장되어 있는 기본식 사전을 이용하여 형태소 분석을 수행하는 형태소 분석부(10)와, 상기 형태소 분석부(10)에 의해서 계산된 형태소 분석 대상 어절의 모든 후보의 점수가 추정 임계치 g 이하이면, 미등록 명사가 포함된 추정 대상 어절이라고 판단하고 추정 대상 어절을 어절 길이의 경우의 수만큼 좌측부분과 우측부분으로 나누어 각 좌측부분 및 우측부분에 대하여 미등록 명사 추정 후보의 점수를 계산하며, 가장 높은 점수의 미등록 명사 추정 후보를 미등록 명사로 추정하고 추정된 미등록 명사를 수집하는 미등록 명사 추정/수집부(40)와, 형태소 분석 대상 어절에 대해 분석된 형태소를 태깅하고 상기 추정 대상 어절에 대해 추정된 미등록 명사를 태깅하는 태깅부(20)를 포함한다.

[0044] 또한, 추정 대상 어절을 출력하거나 미등록 명사 추정 후보를 출력하는 디스플레이부(50)와, 상기 출력된 추정 대상 어절을 확인한 사용자로부터 미등록 명사 여부를 결정하는 결정 신호를 입력받는 입력부(60)를 더 포함하여 구성된다.

- [0045] 상기 미등록 명사 추정/수집부(40)는, 미등록 명사 추정 후보의 가장 높은 점수가 반자동 임계치 이하이면 추정 대상 어절의 출력을 결정하고 입력부(60)로부터 입력된 사용자의 결정 신호가 미등록 명사를 나타내는 경우에 상기 추정 대상 어절이 미등록 명사를 포함하고 있음을 결정한다. 상기 미등록 명사 추정/수집부(40)는 입력부(60)를 통해 수신된 사용자의 결정 신호에 의하여 추정 대상 어절이 미등록 명사를 포함하고 있다고 판단될 경우에는 그 미등록 명사를 수집할 수 있다.
- [0046] 추정 임계치 g 의 값을 조정하여 미등록 명사를 포함하는 어절들에 대하여 미등록 명사 추정의 시도의 수위 조절이 가능할 수 있다. 또한 반자동 임계치의 값을 조정하여 미등록 명사 여부의 결정을 사용자에게 의한 반자동의 방법으로 정확하게 태깅할 수 있으므로 훌륭한 반자동 수집기로 활용될 수 있다.
- [0047] 도 2는 본 발명의 다른 실시예에 의한 형태소 분석기를 위한 미등록 명사 추정 방법을 나타내는 흐름도이다.
- [0048] 본 발명에 의한 형태소 분석기를 위한 미등록 명사 추정 방법은, 형태소 분석 대상 어절의 모든 후보의 점수가 추정 임계치 g 이하이면, 형태소 분석 대상 어절이 미등록 명사가 포함된 추정 대상 어절이라고 판단하여 미등록 명사 추정을 시도하는 제 1 단계(S12), 추정 대상 어절을 어절 길이의 경우의 수만큼 좌측부분과 우측부분으로 나누어 각 좌측부분 및 우측부분에 대하여 미등록 명사 추정 후보의 점수를 계산하는 제 2 단계(S14, S16, S18, S20, S22, S24, S26), 및 가장 높은 점수의 미등록 명사 추정 후보를 미등록 명사로 추정하는 제 3 단계;를 포함한다.
- [0049] 상기와 같은 본 발명에 의한 형태소 분석기를 위한 미등록 명사 추정 장치의 동작을 설명하기로 한다.
- [0050] 미등록 명사 추정/수집부(40)는 형태소 분석부(10)에 의해서 계산된 형태소 분석 대상 어절의 모든 후보의 점수가 추정 임계치 g 이하이면, 미등록 명사가 포함된 추정 대상 어절이라고 판단하여 미등록 명사 추정을 시도한다(S12).
- [0051] 미등록 명사 추정/수집부(40)는 추정 대상 어절을 좌측부분L과 우측부분R 두 부분으로 나눈다(S14). 이 때 나누는 방법은 어절의 길이만큼 경우의 수가 존재하기 때문에 순환 처리를 한다. 즉, 좌측부분이 1개인 경우부터 순차적으로 반복처리하여 어절의 길이 N개의 경우까지 처리한다.
- [0052] 좌측부분 L이 결정되면 어절의 나머지는 우측부분 R이 되고, 우측부분R은 대부분 조사부분에 해당한다. 단, 여기서 구체적으로 R이 반드시 조사여야 한다고 엄격하게 제한하지 않는다. R은 등록된 명사가 될 수도 있으며 ‘이/VCP’로 시작할 수도 있다. 여기서 중요한 것은 R이 어떤 형태로든 L과 조합이 가능하다면 미등록 명사 추정 후보로 생성될 수 있어야 한다는 것이다. 결국은 본 발명이 적용되는 형태소 분석기(UTagger)의 후보 점수 계산에서 L과 R의 조합 가능성을 계산할 것이기 때문에 R이 기분석 사전에 존재하는지를 먼저 판단한다.
- [0053] 길이 1의 좌측부분L에 대응되는 우측부분R이 기분석 사전에 존재하지 않으면, S24 단계로 이동하여 길이 2의 좌측부분L 및 우측부분R의 경우로 넘어가서 다시 우측부분R이 기분석 사전에 존재하는 지를 확인한다(SS14, S16).
- [0054] 우측부분R이 기분석 사전에 존재하면, 기분석 사전에 존재하는 우측부분의 길이를 획득하고, 획득된 우측부분의 길이 및 상기 좌측부분의 길이를 [수학식 1]에 입력 연산하여 상기 좌측부분의 빈도 점수를 계산한다(S18).
- [0055] 또한, 기분석 사전에 존재하는 우측부분의 빈도를 [수학식 2]를 이용하여 수정한다.
- [0056] 그런 다음, 미등록 명사 추정/수집부(40)는 [수학식 1]을 이용하여 계산된 좌측부분의 빈도 점수, 기분석 사전에 존재하는 우측부분의 빈도로부터 수정된 빈도값, 상기 좌측부분 및 상기 우측부분의 조합 가능성 정도를 곱하여 미등록 명사 추정 후보의 점수를 계산하고(S20), 미등록 명사 추정 후보를 디스플레이부(50)에 출력할 수 있다(S22).
- [0057] 이와 같은 순환처리가 완료되면, 미등록 명사 추정/수집부(40)는 가장 높은 점수의 미등록 명사 추정 후보를 미등록 명사가 포함된 어절로 추정한다. 미등록 명사 추정/수집부(40)는 점수가 지나치게 낮은 미등록 명사 추정 후보들을 삭제할 수도 있다.
- [0058] 본 발명에 의한 형태소 분석기를 위한 미등록 명사 추정 방법에 의하여 추정 대상 어절 중에서 미등록 명사로 정상적으로 추정된 어절들은 "바잉오피스, 핸드캡, 컬럼니스트, 썬키스트, 플룸, 잔탁의, 리네이취링, 여세부쟁, 비스므쓰, 스트립댄서, 얼짱, 퍼블리셔스" 등등이 있다.
- [0059] 반복 실험 결과, [수학식 1] 및 [수학식 2]는 [수학식 3] 및 [수학식 4]와 같이 설정될 수 있다. 이는 하나의

예시에 불과하며 이에 한정되는 것은 아니다.

수학식 3

$$l = \frac{((\frac{1}{16}) \times s_r + 0.5)}{200^{2 \times s_r - 5}}$$

[0060]

수학식 4

$$r' = r^{0.075}$$

[0061]

[0062] 본 발명이 적용되는 형태소 분석기(UTagger)의 경우, 세종말뭉치에서 어절 전체가 1번이라도 등장할 경우에 점수 대신에 기본값 사전에서 검색된 빈도가 적용된다. 추정 임계치 g의 값은 1로 설정될 수 있다.

[0063] 한편, 미등록 명사 추정/수집부(40)는 미등록 명사 추정 후보들의 가장 높은 점수가 반자동 임계치 이하이면 사용자의 확인이 필요한 경우라고 판단하여 상기 추정 대상 어절을 출력한다. 따라서, 출력된 추정 대상 어절을 확인한 사용자로부터 미등록 명사 여부를 결정하는 결정 신호가 입력부(60)를 통해 입력되면, 미등록 명사 추정/수집부(40)는 상기 추정 대상 어절이 미등록 명사를 포함한 어절이라고 판단하는 것이다.

[0064] 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 따라서, 이상에서 기술한 실시예는 모든 면에서 예시적인 것이며 한정적인 것이 아닌 것으로 이해해야만 한다. 그러므로 본 발명의 진정한 기술적 보호범위는 아래의 특허청구범위에 의하여 정해져야 할 것이다.

부호의 설명

[0065]

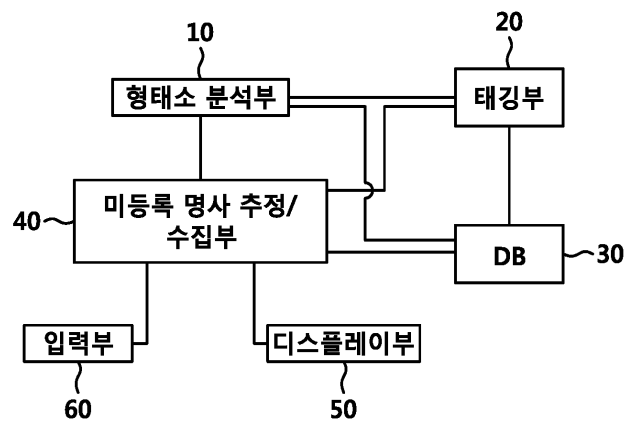
10: 형태소 분석부 20: 태깅부

30: DB 40: 미등록 명사 추정/수집부

50: 디스플레이부 60: 입력부

도면

도면1



도면2

